Contemporary Advances in Physics—XV

The Classical Theory of Light, First Part

By KARL K. DARROW

FOR twenty years and more we have been hearing continually about the conflict between the corpuscular and the undulatory theories of light, and it is possible that for years to come we may be hearing about a similar contest between the wave-theory and the particle-theory of matter. Furthermore, there are intimations that if an adequate theory either of light or of matter ever is attained, it will involve conceptions of waves which in certain limiting cases approach to conceptions of particles. Already it is established that the appropriate way to attack the typical problems of the atom consists in setting up a wave-equation, and dealing with it in the same manner as one adopts to solve the typical problem of acoustics: how to determine the resonance-frequencies of a piece of elastic matter, such as a taut wire or a drumhead or a column of air in a tube. Therefore it seems opportune to restudy, with care and in detail, the great classical example of a wave-theory highly developed and widely successful—the great theory of light dimly foreshadowed by Huygens, endowed with its essential attributes by Young and Fresnel and Kirchhoff and a host of their coevals, utilized in the design of a multitude of ingenious instruments, perfected by Maxwell and connected with the theory of electricity and magnetism, and serving to this day as the basis for the theory of quanta. So doing, we shall be reminded of many triumphs of the past century of physical research, discoveries which in their time were as exciting as new quantum phenomena in ours; we shall notice certain achievements themselves as recent as those of quanta, and perhaps not less impressive; we shall retrace the reasonings which led to certain conclusions which the quantum-theories, unable to do without them and yet incompetent to derive them, have taken bodily over from their forerunner; we shall reconsider the evidence which in the litigation of a century ago caused the verdict to be rendered in favour of the wave-theory over the particle-theory; and perhaps incidentally we shall be drilling ourselves to test the evidence lately submitted and still to be submitted in the appeal of that case, and in the hearing of that other which impends.

For purposes of drill it might seem better to study the example of water-waves, which are visible; or sound-waves, of which no one denies

the existence, and no one wishes to supplant them with quanta. Ripples on the surface of a pond do furnish a precious example of wave-motion, and I presume that the notion of an undulatory theory was suggested originally by these; but it is precisely because they are visible that they fail to pose some of the questions which in dealing with light and matter are the most perplexing. Watching the leaves and the straws which float upon the surface of the water, one sees that they do not advance with the ripples; they are heaved up and down as the crests and the troughs of the wavelets pass them by. It is evident that the waves are not to be identified with the water; rather they are a form, a profile, a molding of the surface, which moves rapidly along while the substance of the liquid oscillates only a little. Now in this instance of the ripples on the pond, it is the relatively-immobile water which seems substantial and real, while that which is propagated as a wave appears to be merely a shape or a configuration, nothing more than a geometrical abstraction. It would seem strange and whimsical to assert that the liquid is a mere abstraction, but the waves are real. Yet we have to embrace this apparent absurdity in dealing with waves of light.

The example of sound-waves shows forth the paradox quite clearly. One can feel a tuning-fork; when it begins to act as a source of sound, one can see that it is quivering, and with a stroboscope one can even follow the actual course of its motion; it is even possible to see condensations and rarefactions travelling through the air, and there are numberless indirect ways of showing that sounding bodies and sound-transmitting media are matter in vibration. To the eye and to the hand, the body which vibrates is material and substantial, but not so its "vibration"—this word is only a way of saying that the shape, or the position, or the density of the body is undergoing a continuous and cyclic change.

It happens, however, that we also possess a sense for which the vibration is real but the vibrating substance is not. The ear takes no cognizance of the steel of which the tuning-fork is made, nor of the air which carries the undulations; but the ear perceives a tone. One must fully realize that the sense of hearing does not disclose that sound is vibratory. The ear does not report a sensation which goes through a cyclic variation two hundred and fifty-six times (or whatever the frequency of the fork may be) in every second. If it did, it would be perceiving the vibrating medium, not the vibration. The ear reports a sensation which is uniform, unvarying, constant; in fact, it translates a steady vibration into a constant sensation. This we have learned with ease, because of the collaboration of the other senses

which observe the bodies which oscillate. But if no one had ever felt or seen the quiverings of the humming fork, the ringing bell, or the resounding drumhead, we should be handicapped severely for discovering the true nature of sound.

Precisely so handicapped are we for discovering the nature of light. It may be that light is the vibration of a substance; but if so, the eye does not perceive that substance nor anything which fluctuates; it translates the vibration into a constant sensation. Moreover, we have no other sense which perceives that substance. When the filament of a lamp is incandescent, nothing is observed to pulsate on its surface; nothing is observed to go up and down or back and forth in the surrounding vacuum. Our instruments also fail to detect anything of which the vibrations are light. One may measure light with a photographic film or a bolometer; but the undulations—if such there be—are translated, in the one case into a steady rate of chemical change, in the other into a steady flow of electric current. In short, the eye and all our instruments register light as the ear registers tone, and not at all as the eye or the hand may register the quiverings of a sounding body; and therefore they do not report that light is vibratory. And if it be true that tangible matter is itself of the nature of a wave-motion, then the sense of touch must respond to these waves as the eye to light or the ear to sound, not reporting anything vibratory and not perceiving any medium which vibrates, but translating the vibrations into a constant sensation.

Therefore, to test whether light or matter or electricity is a wave-motion, one must make such experiments as could be made to test whether sound is a wave-motion, if there were no instrument able to perceive the vibrations of matter except the ear. Let us then suppose ourselves required to prove, to someone unable or unwilling to use any instrument except the ear, that sound is of the nature of waves; and consider how we should go about it.

The ear, we are told, is able to make distinctions of "loudness," "timbre," and "pitch." The two latter, interesting as they are, are of no immediate concern in this enterprise. It is sufficient to know that the ear makes distinctions in loudness, which according to the wave-theory correspond to distinctions in amplitude of vibration. Again, we have no concern with the exact relation between amplitude and loudness. What matters is that the former controls the latter, and therefore the latter reveals the former. Though the ear cannot detect the cyclic variations of the density and pressure of the air which make the sound, it can detect fluctuations in the amplitude of these cyclic variations. To put this statement into briefer language of

which, much later, there will be a reminiscence: the ear can detect the amplitude, but not the phase, of the vibrations. It follows, then, that we must devise tests of the wave-theory in which the amplitude of the waves shall vary from time to time, or from place to place.

Such tests are easily arranged. Let a pair of tuning-forks be set up not too close together and not too far apart. If sound consists of waves, the spherical undulations broadening outward from each of these separately must fall off steadily in amplitude as they recede, and the sound grow steadily fainter as the listener moves away. So it does; but the fact is equivocal, and cannot be taken as evidence for the waves; if the fork emitted corpuscles of sound, they would scatter apart as they flew away, and fewer would enter the ear the farther off it was placed. If, however, both of the forks are giving voice at once, and trains of spherical waves expanding outward from both, then the amplitude in the air must vary in a curious and striking manner from place to place, alternating between maxima and minima. This is just the sort of test which the ear is excellently fitted to make; being moved (or the mouth of its listening-tube being moved) from place to place in the field where the streams of sound overlap, it reports the fluctuations of loudness which are predicted from the wave-theory. By properly choosing the conditions one or more of the minima may be reduced to zero; loudness added to loudness makes silence. By properly choosing the conditions, maxima and minima may be caused to move in succession across a fixed point, listening whereat the observer hears "beats." All of these are phenomena of *interference*, and many like them are realized with light.

But it is not necessary to produce two overlapping streams of sound, in order to find evidence favouring the wave-theory. One suffices, provided that we try to separate from it a narrow jet or ray. Near one of the forks let a wall be placed, and perforated with a little hole. This seems to be an artifice for producing a constricted beam of sound proceeding like a searchlight straight outward along the line passing from the source through the hole; but it does not work that way. Instead, the tone of the fork is heard everywhere beyond the wall; sound is radiated from the hole in all directions. The aperture becomes itself a sort of secondary source, from which sound emanates sidewise as well as forward.

Precisely similar is the visible behaviour of water-waves (and incidentally of the violent compressional pulses produced in air by explosions, which have been photographed; but we are assuming that our imaginary pupil knows nothing of sound but what he hears!). Circular ripples expand over the surface of still water until one of them

meets a wall with a very narrow opening. Does a narrow segment of the ripple go clean through the opening and continue onward as a sharply-ended crescent? Not so; a new circular or semicircular ripple spreads out from the aperture as a new centre.

This is called, in the science of light, a phenomenon of *diffraction*. Actually, it is a phenomenon which reveals the law of wave-propagation—a law, which in the deceptively simple cases of spherical, circular, or infinite plane waves is artfully concealed. When one sees a circular ripple broadening over the surface of a lagoon, it seems as if each arc of the circular crest were advancing independently and of its own momentum; as if each segment of the circle at a given moment were due entirely to the corresponding segment of the smaller circle which existed a fraction of a second earlier. Nothing could be further from the truth. At a given moment, a given segment of the circle is due to the collaboration of all the segments of the earlier smaller circle; and it will collaborate with all the segments of its own circle to build the future yet larger one; and if isolated from the rest of the circumference, it would build a new family of circular ripples all by itself. Somewhat as the primitive animals which can regenerate their amputated parts, a wave-system seems to possess in each of its elements something of the power to build itself anew.

Such is the nature of ripples on water and sound-waves in air. As for a general definition of wave-motion, perhaps there is no better way of making one than to accept this manner of propagation as the distinctive mark. It may seem strange, however, that there should be any question about definition. Does not everyone know what a wave is? and is not the difference between a wave-theory and a corpuscle-theory made instantly clear by their names?

Well! it would not be hard to compile a series of paradoxical statements, by which to show that our immediate off-hand notion of a "wave" is not by any means sufficiently precise to serve as basis for an elaborate physical theory. Even in the ancient and familiar instance of circular ripples on water, even for students acquainted with the concepts of wave-length and wave-speed, there are possibilities of confusion. It is not expedient to define the wave-length as the distance from one crest to the next, for this is inconstant. It is not expedient to define the wave-speed as the speed with which a crest advances, for this may depend upon the form of the wave. It is injudicious to think exclusively about the profile of the water-surface as a sequence of visible elevations and depressions gliding steadily onward without change of shape; for any part of the profile may alter itself incessantly as it advances, departing more and more

from its original contour till it becomes unrecognizable. Definite as the wave-length and the wave-speed and the waves themselves may seem at times to be, at other times they seem indefinite and indefinable.

Now in the general theory these difficulties are removed, for the attention is focussed first of all upon an abstract entity with a neutral name—the "phase." There is a differential equation, a "wave-equation," governing the phase; and this entity is propagated in a certain very definite way conforming to the vague description which I gave above, and it has a wave-length and a wave-speed. As for the elevations and depressions of the water-surface, they copy the variations of the phase more or less faithfully, and may be computed from these; but except in particular cases the copy is not exact. To the theorist, the ripples upon the water appear as the secondary and imperfect manifestations of an abstract wave-motion, discernible only to the eye of the mind.

That the undulatory theory should introduce an abstraction even into the example whence it sprang, requiring us to imagine waves of phase underlying the tangible waves of the sea, is not at all remarkable. Often in physics a theory evolves in this way. It begins when some one notes a resemblance between two or more phenomena; it continues by the invention of a neutral and colorless mathematical expression for describing the common aspect of all these phenomena; and then the theorists take over the mathematical expression and transform and generalize and extend it, until the theory, which at first was a casual statement that two different things are in some ways much alike, eventually is defined as the entire system of solutions of some differential equation. At present there seems to be no adequate way of defining the term "wave-theory," except to say that wherever a certain differential equation is introduced and solved, there a wave-theory is adopted. Moreover, it is open to anyone to adduce new differential equations more or less like the first one, and define as a wave-theory any which involves the solutions of these. Then a wave-motion is any motion which conforms to one of the solutions; and when it comes to defining a wave, what can anyone say except that under certain restricted conditions a wave-motion may resemble a procession of ripples on water?

Such a consummation may be devoutly wished by the mathematician; to the physicist and to the expositor it is not always so welcome. As a theory increases in scope through increase of abstraction, it loses the picturesqueness which for many minds is its reason for being. One climbs and climbs, and the view indeed grows wider, but the

fascinating details of the landscape are distorted when seen from above, and finally they are lost in the haze of distance. It grows more difficult to lead others to the heights, and sometimes even the explorer cannot retrace his path and return to the firm ground of experience whence he departed. Yet repeatedly in the evolution of physics it happens that a theory, already grown so abstract that it seems almost completely severed from reality, suddenly makes new contact with the world of phenomena by a prediction so novel and daring that except for the far preliminary excursion it would probably never have been conceived; as for instance the existence of quanta, the "Einstein shift" of the lines in the spectrum of the sun, the diffraction of electrons by crystals. Remembrance of such episodes as these is an encouragement, when the path seems devious and steep.

Propagation of Waves

The laws of the propagation of waves—the so-called "laws of diffraction"—are the most important topic with which we have to deal; for they involve the very nature and definition of wave-motion, and in the end the distinction between a corpuscular and an undulatory theory of matter may rest upon these. Let us attend first of all to the making of this distinction.

Imagine, then, a multitude of particles—bullets, or atoms, or sand-grains—all rushing along through space in the same direction with the same speed, say northward with the speed c. Suppose that the location of each is stated for a certain moment of time, say t_0. The question to be asked is a very simple one, of the yes-or-no variety. At an arbitrarily-chosen point P, at an arbitrarily-chosen moment t, will there be a grain of sand or will there not?

It is easy to see what determines the answer. If at the point P at the moment t there is a grain of sand, it must have spent the time-interval extending from t_0 to t in travelling northward along the straight south-to-north path which ends at P, and therefore commences at a point P_0 due south of P and distant from it by $c(t - t_0)$ units of length. If therefore at the moment t_0 there is a grain of sand at P_0, the answer to the question is *yes*. Otherwise it is *no*. No other knowledge is required, or even relevant. It is not necessary to know the location of any of the particles which are not upon the north-south line traversing P. It is not even necessary to know the location of any particle which is upon that line, provided that at the instant t_0 it is surely somewhere else than at P_0. The state of affairs in P at t is controlled by the state of affairs in P_0 at t_0, and by nothing else whatever.

Let the same question be put in another way. Be it supposed that we are required to predict whether or not there will be a particle in the place P at the moment t; and that we are offered our choice of data concerning the places of the particles at any prior moment. Let us choose at random some point P' due south of P, distant from it by r' units of length. Then there is only one piece of information for which we have to ask: is there a sandgrain passing through P' at the moment $(t - r'/c)$, earlier than t by r'/c units of time? Any other information would be not only superfluous, but useless. Had we chosen some point lying south of P at a distance r'', the condition prevailing there at the moment $(t - r'/c)$ would have had no bearing upon the problem; but the condition there at the moment $(t - r''/c)$ would have been all-powerful. Had we chosen some point not lying south of P, nothing happening there at any time would have had any bearing upon the question to be answered.

In fine: at any moment t' there is a corresponding point P' lying south of P, which holds the destiny of the point P at the moment t. That which is predestined to befall in P at t is at every prior instant concentrated, so to speak, at a particular point of space. As time draws on towards t, this point moves on toward P, travelling always along the north-south line—travelling always, let us say, along a certain *ray* which at the proper moment carries it right into P.

All these remarks may seem too evident and trivial to be worth the making; yet they deserve attention, for it is here that the contrast lies between motion of particles and motion of waves, between undulatory theories and corpuscular. If the region around the point P is traversed not by corpuscles but by waves, it is not correct to say that the condition at the point P at the moment t is determined by the condition in some other *point* at some prior moment. Even if the waves appear to be travelling northward with the constant speed c, it is not right to say that the state of affairs prevailing in P at t is controlled entirely by the state of affairs prevailing at the moment $(t - r/c)$ at the point r units southward from P. The destiny of P at t is not travelling towards it concentrated into a point moving along a ray. Under some circumstances it appears to be right to say so; but this is only a semblance, as experiments in other conditions will clearly prove.

Suppose for instance that one is confronted with the task of sheltering the point P, first against corpuscles and then against waves, which are advancing from the south. It seems natural to put some obstacle athwart that particular north-south line which traverses P; for example, to place a solid disc so that its axis lies upon that line. If the disc can arrest all the particles which fly towards it, and cannot deflect

those which do not, then no matter how small it is it shields the point P completely against corpuscles. Not, however, against waves. Suppose that the point P is in water, where the actual waves may be seen; suppose that before the obstacle is dipped in, each of the wave-crests extends straight east and west, and they move straight northward. When the obstacle is inserted southward from P, the water at P does not become perfectly quiet. Apparently the waves curl around the edge of the obstacle, invading the zone behind it which it could have protected perfectly against corpuscles. One cannot stop a wave-motion from reaching a point merely by interrupting with some small obstruction the line along which the waves seem to be approaching it.

Now what this means is simply that, whether the obstacle be present or absent, and even though the undisturbed wave-crests move steadily due northward, the motion at P is not controlled exclusively by the motions at earlier moments at the points due south of P. To put it a little more loosely: the wave-motion at a point arrives not solely from the direction from which the wave-fronts appear to be coming, but from all directions. To put it much more strictly: imagine a sphere drawn, with any radius r, around P as centre. When we were dealing with corpuscles, we found that the state of affairs at the centre of this sphere at the moment t was entirely controlled by the state of affairs at the moment $(t - r/c)$ at one single point on the sphere (the point due south from P). Now that we are dealing with waves, we shall find that the state of affairs at the centre of the sphere at t depends upon the state of affairs all over the sphere at $(t - r/c)$. Every point upon the sphere influences the centre. Every point in the medium which the waves traverse sends forth an influence to every other point; the influence is not instantaneous, but travels from one point to another with the wave-speed c. This "influence" is often called the *wavelet.*

Too much emphasis has been laid, in the foregoing passage, upon the spheres which are centred at P; and this must now be rectified. Any closed surface whatever may be drawn around P, and the state of affairs in P at t will be determined by the state of affairs prevailing all over this surface, S, at certain prior moments t'; only, since the area-elements of S are not in general equidistant from P, the corresponding values of t' are not in general the same for all of them. The distance r, measured from P along any direction to the surface S, is in general a function of direction; consequently the time-interval, r/c, required for a wavelet to arrive at P from S along any direction is itself a function of direction; and so also is t', which is $(t - r/c)$. To every point P' on S corresponds its own value of t'; and if we know the wave-motion in every P' at its proper t', we can determine the wave-motion

in P at t. Therefore, if there is any closed surface in space, everywhere over which the wave-motion is known for all times, it is possible to compute the wave-motion at any point in the volume which that surface encloses.*

This is a feature common to all the familiar examples of wave-motion, and it is suitable for a tentative basis for a general definition of waves.

To formulate it strictly, let s be used as the symbol for any quantity which is propagated in waves. Examples of such a quantity are: the twist of a taut and twisted wire—the lateral displacement of a taut wire or a tense membrane—the excess of the pressure in the air over its average value—a component of the electric field-strength or the magnetic field-strength in a vacuum—the entirely imperceptible and hypothetical entity denoted by Ψ in wave-mechanics.

We write s as a function of x, y, z, and t:

$$s = s(x, y, z, t). \tag{1}$$

Fewer than three dimensions of space will suffice in some cases (e.g., those of the wire and the membrane); in certain problems of wave-mechanics, more than three may be required; but in dealing with sound in air and light in vacuo, three are usually necessary and sufficient. For the time being I will suppose that the speed of the waves is everywhere the same. Interesting things will happen when this assumption is discarded.

What I have loosely called "the state of affairs" in a point $P(x, y, z)$ at a moment t will involve the value of s at x, y, z, and t. Also it may involve the first and higher derivatives of s with respect to space and time, evaluated at x, y, z, t. Which of these derivatives we are required to know is something which might vary from case to case. For the present, we may consider ourselves required to know s and its first derivatives ds/dx, ds/dy, ds/dz, ds/dt.

We are to evaluate s and its derivatives at a point P at a moment t, in terms of the values which s and its derivatives possessed at certain earlier moments over a surface S enveloping P.

Let P be made the origin of our coordinate-system; let x, y, z denote the coordinates of the points on the surface S; let r denote the distance from the origin to any of these points, so that:

$$r^2 = x^2 + y^2 + z^2. \tag{2}$$

Introduce as an auxiliary the function U, defined thus:

* Naturally the surface must not be so drawn that it includes sources emitting waves during the time-interval $(t' - t)$.

$$U(x, y, z, t) = s(x, y, z, t - r/c). \tag{3}$$

The value of U in any point of the surface S at the moment t is the value of s which prevailed in that point at the moment when the "wavelet" started forth which was destined to reach the origin at t. It might be said that an observer, stationed in the origin at the moment t and inspecting the surface by means of the wavelets, observes the values of U instead of the contemporary values of s. Thus a star-gazer viewing the sky perceives, not the stars as they now are or as at some one past moment they all were, but each star separately as it was at some past epoch peculiar to itself; and the apparent arrangement of the heavenly bodies is one which in fact has never existed.

We shall be concerned not only with the value of s, but with the values of the space-derivatives ds/dx, ds/dy, ds/dz, which prevail at each point of the surface at the moment when the wavelet starts forth; for all of these will influence the value of s at the origin when the wavelet arrives there. These may be written as derivatives of U; but one must be careful here, for U is a function of x, y, z not only explicitly, but also implicitly through r; and there is a distinction to be made between total and partial derivatives, a distinction having physical importance.

To grasp this, denote by (x, y, z) the coordinates of some particular point on S, and by $(x + dx, y, z)$ those of a nearby point, and by r and $r + dr$ their respective distances from the origin, and by U and $U + dU$ the values of U in these points at the instant t. Now, U and $U + dU$ are values of s which existed *at different instants of time*, as may be seen by writing down the expressions:

$$U = s(x, y, z, t - r/c); \quad U + dU = s(x + dx, y, z, t - \overline{r + dr}/c). \tag{4}$$

Therefore, if I form the total derivative dU/dx in the classical way, I am *not* obtaining the value of ds/dx which prevailed in (x, y, z) at $(t - r/c)$. To obtain this value, I must begin by subtracting the value of s prevailing in (x, y, z) at $(t - r/c)$ from the value of s prevailing in $(x + dx, y, z)$ at the same moment; that is, I must form the difference between $(U + \partial U)$ and U, meaning by the former symbol:

$$U + \partial U = s(x + dx, y, z, t - r/c). \tag{5}$$

I must then divide this difference by dx, and pass to the limit. But this is the classical way of forming the partial derivative of U with respect to x. Therefore the values of the derivatives ds/dx, ds/dy, ds/dz prevailing at the moment of departure of the wavelet which is destined to reach the origin at t, are the partial derivatives $\partial U/\partial x$,

$\partial U/\partial y$, $\partial U/\partial z$. However, the value of the derivative ds/dt prevailing at the moment when the wavelet starts is simply the derivative dU/dt, which we may as well write $\partial U/\partial t$—it makes no difference.

Our definition of wave-motion may now be stated more rigorously. A quantity s is said to be propagated by waves, if its value at the origin at the moment t is determined by the values of U, $\partial U/\partial x$, $\partial U/\partial y$, and $\partial U/\partial z$ over any surface enveloping the origin.

We now turn to another and more familiar definition of wave-motion, which shall presently be shown to fall as a special case under this one.

The Wave-Equation

There is a very celebrated differential equation of mathematical physics, known as "the wave-equation" *par excellence*. Any theory which culminates in this equation is designated as a wave-theory. The foundation of the theory of sound is the proof that the excess of pressure in the air over its average value is subject to this equation. The elastic-solid model of the luminiferous æther was partially suited to explain the phenomena of light, because the compressions and the distortions of an elastic solid conform to the wave-equation. The electromagnetic theory of light was born when Maxwell discovered interrelations between electric and magnetic fields, out of which by transformation a wave-equation could be formed. Undulatory mechanics is based upon an equation of this type which emerges during the process of setting and solving the classical equations of motion.

This wave-equation is:

$$c^2\left(\frac{d^2s}{dx^2}+\frac{d^2s}{dy^2}+\frac{d^2s}{dz^2}\right)=\frac{d^2s}{dt^2}\cdot \tag{6}$$

To demonstrate why it is called a wave-equation and what is the physical meaning of the constant c, it is customary to make a drastic simplification by assuming that the function s depends only on one coordinate. Such is the case, for instance, when s stands for the transverse displacement of an endlessly long taut string initially parallel to the axis of x; likewise, when it stands for the excess of the pressure of the air over its average value, and this excess is constant over every plate normal to the x-direction—a condition known as that of "plane waves." Then the wave-equation assumes the form:

$$c^2\frac{d^2s}{dx^2}=\frac{d^2s}{dt^2}\cdot \tag{7}$$

There are infinitely many solutions of this equation, and among them

are all [1] the functions of the pair of variables x and t, in which these variables appear coupled together into the linear combination $(x - ct.)$ Using f as the general symbol for a function, we may write

$$s = f(x - ct). \tag{8}$$

When such a relation prevails, any value of s which occurs at a given place x at a given moment t recurs at any other moment t' at another place x', distant from x by the length $(t' - t)/c$. *All* of the values of s existing at t are found again in the same order at t', but they have all glided along the x-direction through the same distance $(t' - t)/c$. The form, the profile, the configuration of the string are moving along with the speed c, although the substance of the string is oscillating only a little, and not even parallel to the x-direction. Now this is the property which to a certain degree of approximation ripples on water display; this in fact supplies the elementary and restricted definition of wave-motion, out of which by generalization and extension the wave-theory has grown.

Thus we see that there is reason for calling (7) a wave-equation, and identifying the constant c with the speed of the waves. Yet there are also solutions of (7) which are not of the form (8), and these do not correspond to an unchanging profile of the string travelling along with a constant speed, though by mathematical artifice they may be expressed as a summation of such; and nothing is easier than to find solutions in two or three dimensions of the general equation (6) which do not bear the least resemblance to a regular procession of converging, flat, or diverging waves. The question then arises: is there a feature common to all solutions of the "wave-equation" fitted to serve for a general definition of wave-motion?

I will now show—in the manner of Kirchhoff and Voigt—that there is such a feature, and it is precisely the one already proposed as a definition for wave-motion. If s be a function conforming to (6), and U a function related to s according to (3), then the value of s at any point at any moment is determined by the values of U and its partial derivatives $\partial U/\partial x$, $\partial U/\partial y$, $\partial U/\partial z$, over any surface surrounding that point.* The proof is long and intricate; bur for anyone who desires appreciate the nature of wave-motion, it is not superfluous.

To prove the theorem we have to manipulate the vector (call it W) of which the components are:

[1] Exceptions being made for functions which do not have derivatives, and other curiosities of the mathematicians' museum.

* The necessary requirements for continuity in s exclude sources of light from the region of integration.

$$W_x = \frac{1}{r}\frac{\partial U}{\partial x}, \qquad W_y = \frac{1}{r}\frac{\partial U}{\partial y}, \qquad W_z = \frac{1}{r}\frac{\partial U}{\partial z}. \tag{9}$$

Like U, it is a function of x, y, z not only explicitly, but also implicitly through r; we must therefore discriminate with care between total and partial derivatives. For reference, here are the formulæ[2] connecting derivatives of the one type with those of the other:

$$\frac{d}{dx} = \frac{\partial}{\partial x} + \frac{\partial r}{\partial x}\frac{\partial}{\partial r} = \frac{\partial}{\partial x} + \frac{x}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial x} + \cos(x, r)\frac{\partial}{\partial r}, \tag{10 a}$$

$$\frac{d}{dy} = \frac{\partial}{\partial y} + \frac{\partial r}{\partial y}\frac{\partial}{\partial r} = \frac{\partial}{\partial y} + \frac{y}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial y} + \cos(y, r)\frac{\partial}{\partial r}, \tag{10 b}$$

$$\frac{d}{dz} = \frac{\partial}{\partial z} + \frac{\partial r}{\partial z}\frac{\partial}{\partial r} = \frac{\partial}{\partial z} + \frac{z}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial z} + \cos(z, r)\frac{\partial}{\partial r}, \tag{10 c}$$

$$\frac{d}{dr} = \frac{\partial}{\partial r} + \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} + \frac{\partial z}{\partial r}\frac{\partial}{\partial z}$$

$$= \frac{\partial}{\partial r} + \cos(r, x)\frac{\partial}{\partial x} + \cos(r, y)\frac{\partial}{\partial y} + \cos(r, z)\frac{\partial}{\partial z}. \tag{10 d}$$

The procedure consists in forming the expression for the true divergence of W, to wit:

$$\operatorname{div} W = \frac{dW_x}{dx} + \frac{dW_y}{dy} + \frac{dW_z}{dz}, \tag{11}$$

and integrating it over the volume comprised between two surfaces: outwardly, the surface S over which the values of U are preassigned, and which envelops the origin at which the value of s is to be computed; and inwardly, an infinitesimal sphere centred at the origin.

It will turn out that the volume-integrals of the various terms either vanish, or else may be converted into area-integrals over the two surfaces. Now the area-integral of any function f over the surface of a sphere of radius R may be written as

$$A = 4\pi R^2 \bar{f}, \tag{12}$$

in which $\bar{f}$ stands for the mean value of f over that surface—a statement

[2] In deriving the first three of these formulæ, use the relation $r^2 = x^2 + y^2 + z^2$ in evaluating $\partial r/\partial x$, $\partial r/\partial y$, $\partial r/\partial z$. In deriving the last, remember that in forming a derivative with respect to r at a point P, the increment dr is always measured along the line extending from the origin through P, for which line $x/r = \cos(x, r) = \text{const.}$; $y/r = \cos(y, r) = \text{const.}$; $z/r = \cos(z, r) = \text{const.}$; hence $\partial x/\partial r = \cos(x, r)$, etc. Or one may arrive by geometrical intuition at the formula,

$$d/dr = \cos(r, x)\,d/dx + \cos(r, y)\,d/dy + \cos(r, z)\,d/dz,$$

from which (10 d) may be obtained by means of (10 a, b, c).

which, of course, is merely the definition of f. The essential thing is that if the sphere is infinitesimal, then in the limit $\bar{f}$ becomes the value f_0 which the function f possesses at the centre of the sphere. If f_0 is finite at the origin, A vanishes in the limit; but if f varies inversely as the square of the distance from the origin, A approaches in the limit a finite value differing from zero. Upon this property our demonstration will depend.

Developing by means of (10 a, b, c) the expression given in (11) for div W, we find:

$$\begin{aligned}\operatorname{div} W = \frac{1}{r}\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}\right)\\ +\frac{1}{r^2}\left(x\frac{\partial^2 U}{\partial x\partial r}+y\frac{\partial^2 U}{\partial y\partial r}+z\frac{\partial^2 U}{\partial z\partial r}\right)\\ -\frac{1}{r^3}\left(x\frac{\partial U}{\partial x}+y\frac{\partial U}{\partial y}+z\frac{\partial U}{\partial z}\right).\end{aligned} \tag{13}$$

The second and third terms on the right may next be beneficially transformed by means of (10 d), using first U and then $\partial U/\partial r$ as the argument of the derivatives in that equation:

$$\frac{x}{r}\frac{\partial U}{\partial x}+\frac{y}{r}\frac{\partial U}{\partial y}+\frac{z}{r}\frac{\partial U}{\partial z}=\frac{dU}{dr}-\frac{\partial U}{\partial r}, \tag{14 a}$$

$$\frac{x}{r}\frac{\partial^2 U}{\partial x\partial r}+\frac{y}{r}\frac{\partial^2 U}{\partial y\partial r}+\frac{z}{r}\frac{\partial^2 U}{\partial z\partial r}=\frac{d}{dr}\frac{\partial U}{\partial r}-\frac{\partial^2 U}{\partial r^2}, \tag{14 b}$$

and so finally we arrive at

$$\operatorname{div} W=\frac{1}{r}\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}-\frac{\partial^2 U}{\partial r^2}\right)+\frac{1}{r^2}\frac{d}{dr}\left(r\frac{\partial U}{\partial r}-U\right) \tag{15}$$

as the expression to be integrated over the volume between S and the infinitesimal sphere.

Now *owing to the nature of the function U, the first term of the expression vanishes.* This is responsible for our theorem; for the volume-integrals of the remaining terms can easily be translated into surface-integrals over S and the infinitesimal sphere, from which it will follow that the value of s at the origin is determined by the values of U and its derivatives over S; but if this first term should remain, its volume-integral could not be thus transformed, and we should find that the value of s at the origin was influenced by the values of U all through the space which S encloses.

That the term in question does actually vanish is easily proved. For on the one hand it follows, from the coupling of t and r into the linear combination $(t - r/c)$ in the argument of U, that

$$\frac{\partial^2 U}{\partial t^2} = c^2 \frac{\partial^2 U}{\partial r^2}, \tag{16}$$

and on the other hand it follows, from the facts that the partial derivatives of U at any point and moment have the same values as the corresponding derivatives of s at the same point at some other moment, while the derivatives of s at *every* point and moment conform to (6)—from these it follows that

$$\frac{\partial^2 U}{\partial t^2} = c^2 \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} \right). \tag{17}$$

Therefore the first term of the right-hand member of (15) is zero everywhere, and we have to perform the volume-integration only over the second:

$$\int \operatorname{div} W dV = \int \frac{1}{r^2} \frac{d}{dr} \left(r \frac{\partial U}{\partial r} - U \right). \tag{18}$$

Employ spherical coordinates for the integration; then the element of volume is $dV = r^2 \sin \theta d\theta d\varphi dr$, and we have:

$$\begin{aligned} \int \operatorname{div} W dV &= \int d\theta \sin \theta \int d\varphi \int dr \frac{d}{dr} \left(r \frac{\partial U}{\partial r} - U \right) \\ &= \int d\theta \sin \theta \int d\varphi \left[\left(r \frac{\partial U}{\partial r} - U \right)_S - \left(r \frac{\partial U}{\partial r} - U \right)_R \right]. \end{aligned} \tag{19}$$

This signifies that the long narrow volume-element comprised within any elementary solid angle $dw = \sin \theta d\theta d\varphi$, and limited at its two ends by the surface of the sphere and the surface S, contributes to the volume-integral the difference between the values of $(r\, \partial U/\partial r - U)$ at its two extremities. Completing the integration by considering all of these volume-elements together, we see that the volume-integral therefore becomes a pair of angle-integrals, those of the function $(r\partial U/\partial r - U)$ over the surface S and over the sphere. We may transform the first of these into an area-integral by reflecting that the elementary solid angle dw intercepts upon the surface S the area-element dS, given by the equation:

$$-dS \cos (n, r) = r^2 dw. \tag{20}$$

in which (n, r) stands for the angle between the normal to dS and the radius r drawn to dS from the origin. We must choose positive

directions for these lines. Let the radius be taken as pointing outward, and the normal as pointing inward towards the volume over which we have integrated. Then the angle is greater than 90° and not greater than 180°, its cosine is negative, and the negative sign must be prefixed to the left-hand member of (20) that dS may be positive. We make this transformation in (19), and the first of the angle-integrals becomes:

$$\int d\theta \sin\theta \int d\varphi \left(r\frac{\partial U}{\partial r} - U \right)_S = \int_S dS \cos(n, r) \left(\frac{1}{r}\frac{\partial U}{\partial r} - \frac{U}{r^2} \right). \quad (21)$$

The second of the angle-integrals in (19) relates to the infinitesimal sphere. We transform it as we did the first. Now, however, the process is more simple, for the radius r is constant and equal to R, and the angle (n, r) is zero; hence

$$dS = R^2 \sin\theta d\theta d\varphi \quad (22)$$

and the second angle-integral becomes:

$$\int d\theta \sin\theta \int d\varphi \left(r\frac{\partial U}{\partial r} - U \right)_R = \int dS \left(\frac{1}{R}\frac{\partial U}{\partial r} - \frac{U}{R^2} \right). \quad (23)$$

Here we meet the situation for which equation (12) was introduced—the integration of a function over an infinitesimal sphere. Denote by f the integrand in (23), viz.,

$$f = \frac{1}{R}\frac{\partial U}{\partial R} - \frac{U}{R^2}, \quad (24)$$

by $\bar{f}$ the mean value of f over the sphere of radius R, by U_0 the value of U at the centre of the sphere, which is the origin. As R approaches zero, the surface-integral in (23) approaches a limit A_0 which coincides with the limit approached by $4\pi R^2 f$:

$$A_0 = \operatorname*{Lim}_{R=0} 4\pi R(\overline{\partial U/\partial R}) - 4\pi U_0. \quad (25)$$

Unless the mean value of $\partial U/\partial R$ should vary as the first or a higher power of $(1/R)$—a possibility which must be guarded against—the first term on the right of (25) will vanish. Under this restriction, then, A_0 is equal to $-4\pi U_0$. Now at the origin U is identical with s, by definition (equation 3). Consequently U_0 is identical with the value of s at the origin at the moment t—the very thing which we set out to calculate. For this—let it be called s_0—we have attained the following equation:

$$4\pi s_0 = \int \operatorname{div} W\, dV + \int dS \cos(n, r)(\partial U/\partial r). \quad (26)$$

We still have a volume-integral in the formula; but there is a very noted theorem whereby it may be transformed with sign reversed into a surface-integral over the two surfaces, S and the sphere, which bound the region of integration. According to Gauss' Theorem, any vector function satisfying certain simple conditions of continuity throughout a region enclosed by a surface enjoys this property: its volume-integral through the region is equal to the area-integral, over the enclosing surface, of the projection of the vector upon the direction of the *outward*-pointing normal. This latter is the same in magnitude and opposite in sign to the projection upon the direction of the *inward*-pointing normal, which it is traditional to prefer. The theorem is not valid, if the vector should exhibit certain singularities within the volume; one of the reasons for introducing the infinitesimal sphere is that the vector W has a singularity at the origin, which point must therefore be excluded from the volume of integration.

Remembering the definition of W (equation 9), we see that its projection upon the direction of the *inward*-pointing normal at any place upon either surface is

$$\begin{aligned} W_n &= W\cos(n, r) = W_x\cos(n, x) + W_y\cos(n, y) + W_z\cos(n, z) \\ &= \frac{1}{r}\left[(\partial U/\partial x)\cos(n, x) + (\partial U/\partial y)\cos(n, y) + (\partial U/\partial z)\cos(n, z)\right] \\ &= \frac{1}{r}(\partial U/\partial n), \end{aligned}$$

in which $(\partial U/\partial n)$ stands for the rate at which the function U, owing to its *explicit* dependence upon x, y, and z, varies as one moves *inward* along the normal to the surface. The distinction drawn in the foregoing pages between partial and total derivatives must be remembered. The partial derivative $\partial U/\partial n$ existing at any point P and moment t is equal to the value which the corresponding derivative ds/dn of the function s possessed at that same point at the earlier moment $(t - r/c)$.

The quantity W_n is to be integrated over the surface of the sphere and over the surface S. However, the integral over the sphere vanishes as the radius of this latter approaches zero, for the same reason—and under the same restriction—as caused the integral of the first term in (25) to vanish. This leaves us with nothing but the integral of W_n over the surface S, so that eventually:

$$\int \text{div } W\, dV = -\int_S \frac{1}{r}(\partial U/\partial n)\, dS. \tag{28}$$

$$4\pi s_0 = \int_S dS\left[\cos(n, r)\frac{\partial}{\partial r}\left(\frac{U}{r}\right) - \frac{1}{r}\frac{\partial U}{\partial n}\right]. \tag{29}$$

We have spoken of the value of s at the origin of coordinates, for mathematical convenience; but in reality the "origin" is any point P, and S is any surface enclosing that point, and s_0 is the value of s in the point P at any moment t, and U is the value of s in any point distant by r from P, evaluated at the moment $(t - r/c)$. Hence (29) may be written thus:

$$4\pi s = \int_S dS \left[\cos(n, r) \frac{\partial}{\partial r} \left(\frac{s(t - r/c)}{r} \right) - \frac{1}{r} \frac{\partial s(t - r/c)}{\partial n} \right]. \quad (30)$$

The task is achieved. It has been proved that when a function conforms to "the wave-equation" it conforms also to the first-suggested definition of a wave-motion, in that its value at any time and place is determined by its anterior values and those of its derivatives over a surface completely enclosing the place. Moreover the actual formula has been derived whereby the value at any point and moment can be computed when the values all over any surrounding surface are known at the appropriate prior moments.

Introduction of the Ideas of Frequency and Wave-length

Hitherto I have spoken chiefly of an extremely abstract "something," denoted by a symbol s, and possessed of the property that its value at any point and moment is built up out of contributions despatched at earlier moments from all of the area-elements of any continuous surface which encloses the point; these contributions being borne as it were by messengers, who travel to the point at a finite speed from the various area-elements whence they depart. Only one physical constant has been introduced, and this is the speed of these messengers. This is the constant which appears in the wave-equation (6), being there denoted by c. It is commonly called the speed of the waves; but, for various reasons which will eventually appear, it had better be called the *phase-speed*. Now there are two other constants familiar in our experience with water-waves and sound; they are *frequency* and *wave-length*. Let us try to import them into the general theory.

At any point of a water-surface over which uniform ripples are passing, the elevation is a periodic function of time; so also are the pressure and the density at any point of a gas through which uniform sound is flowing, or the displacement of either prong of a steadily-humming tuning-fork. Any periodic function of time is either a sine-function, or a composite of sine-functions. It is suitable therefore to begin by analyzing the case in which the function is a sine. Using f to denote any of the quantities above mentioned or anything behaving like them—say displacement, for example—let us write:

$$f = F \sin(nt - \delta) = F \sin \varphi. \quad (41)$$

In this very familiar form, F stands for *amplitude* and n for 2π times the *frequency*, and δ for something which is commonly called the phase; but it will be better to reserve this name for the entire argument of the sine-function:

$$\text{Phase} = \varphi = nt - \delta = \arcsin (f/F), \tag{42}$$

and I shall use it henceforth in this sense.

Now, in general, both the amplitude F and the phase φ vary from point to point—the latter because δ is often a function of position. Consequently f is a function of x, y, z and t; and it is immediately important to find out whether f satisfies the wave-equation. One who is familiar chiefly with the standard one-dimensional case of waves on a string is likely to think that this is true as a matter of course. However, on differentiating f twice with respect to x or y or z and taking due account of the dependence of both F and δ upon these variables, one sees directly that in general it is not true—not unless the functions F and δ conform to definite and sharply restrictive conditions.[3]

This appears a rather disconcerting result. However, if instead of f we envisage f/F—the value of the displacement at each point referred to its amplitude there as a unit, or the sine of the phase—it turns out that the condition under which f/F obeys the wave-equation is far less drastic. Forming the derivatives, we obtain:

$$\frac{\partial^2}{\partial t^2}\frac{f}{F} = -n^2 \sin \varphi, \tag{43}$$

$$\nabla^2 \frac{f}{F} = (\nabla^2 \varphi) \cos \varphi - \left[\left(\frac{\partial \varphi}{\partial x} \right)^2 + \left(\frac{\partial \varphi}{\partial y} \right)^2 + \left(\frac{\partial \varphi}{\partial z} \right)^2 \right] \sin \varphi. \tag{44}$$

Evidently, in order that the function f/F shall conform to the wave-equation, it suffices that

$$\nabla^2 \varphi = 0, \tag{45}$$

This condition is fulfilled by a variety of functions, including all which are linear in x, y, and z—the case of "plane waves," which as we shall see is one of those permissible when the wave-speed is everywhere the same, as I have been assuming.

Next we will evaluate the speed of the phase-waves, and incidentally we shall be led to a new aspect of wave-motion. When the phase

[3] The general expression for $\nabla^2 f$ is $(\nabla^2 F - F|\nabla\phi|^2) \sin \phi + (2\nabla F \cdot \nabla\phi + F\nabla^2\phi) \cos \phi$. The coefficient of $\cos \phi$ must vanish, if f is to satisfy the wave-equation with real phase-speed. By introducing the notion of "imaginary phase-speed" one may continue to regard the function f as conforming to the wave-equation, even though the coefficient of the cos-term does not vanish.

conforms to (45), it follows from (43) and (44) that

$$\nabla^2 \frac{f}{F} = \frac{1}{n^2}\left[\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2 + \left(\frac{\partial\varphi}{\partial z}\right)^2\right]\frac{\partial^2}{\partial t^2}\frac{f}{F}. \tag{46}$$

The quantity in brackets, being the square of the magnitude of the gradient of φ, shall be denoted by the usual symbol $|\nabla\varphi|^2$. Then for the square of the phase-speed we obtain $n^2/|\nabla\varphi|^2$, and for the phase-speed itself:

$$c = = + n/|\nabla\varphi|, \tag{47}$$

The quantity n is 2π times the frequency. Also, it is the time-derivative of φ, as follows directly from the definition in (41); consequently:

$$c = = (\partial\varphi/\partial t)/|\nabla\varphi|. \tag{48}$$

This is an equation with a very important meaning, which will now be displayed.

Select any point in the medium and any moment of time t_0, and denote by φ_0 the value of φ prevailing then and there. Singular cases excepted, there is an entire surface containing the point in question everywhere over which φ has the same value φ_0. This surface is by definition a *wave-front*. Call it S_0. At a slightly later moment $t_0 + dt$, there will also be a surface everywhere over which the value of φ is φ_0. It will however not be the same surface S_0, but another—a wave-front S_1 so placed that from any point P_0 on S_0 the nearest point on S_1 is reached by measuring the length cdt along the line normal to S_0, in the sense in which φ is decreasing (the sense opposed to the gradient of φ).

To see this, imagine a particle which at the instant t_0 is travelling through P_0 along the direction normal to S_0 in the sense just stated, with a speed to be designated by u. At the instant $t_0 + dt$ it occupies a point where the value of φ then prevailing is given by the formula:

$$\begin{aligned}\varphi_0 + d\varphi &= \varphi_0 - |\nabla\varphi|ds + \left(\frac{\partial\varphi}{\partial t}\right)dt \\ &= \varphi_0 - u|\nabla\varphi|dt + \left(\frac{\partial\varphi}{\partial t}\right)dt,\end{aligned} \tag{49}$$

for in the time-interval dt it travels over a distance $ds = = udt$ along the normal to the surface S_0, and along this normal the slope of the function φ is equal to $|\nabla\varphi|$, and meanwhile at each point of space φ is varying directly with time at the rate $(\partial\varphi/\partial t)$. Now if the imagi-

nary particle happens to be moving with just the speed defined by the equation

$$u = (\partial\varphi/\partial t)/\nabla\varphi = c, \tag{50}$$

the coefficient of dt in equation (49) vanishes; that is, the particle as it moves along keeps up with the preassigned value of φ; but this is the same thing as saying that c is the speed of the wave-front.

The phase-function φ therefore possesses a quality which in itself suggests one aspect of a wave-motion. It is not periodic, neither does it conform to the wave-equation; but each of the surfaces over which φ has any constant value is perpetually travelling. Each of them may be changing continually in size, it may even be changing in shape; but each retains its identity, and if it is completely known for any given instant, its past and future history are determined completely; for each of the area-elements of such a surface is moving at the speed c and in the direction normal to itself, and from the position of each area-element at the moment t we can determine the position of the area-element into which it evolves at the moment $t + dt$, and repeat this process of prediction or retrospect *ad infinitum*. I will speak of this state of affairs as *propagation by wave-fronts*.

Having determined by (48) the relation between frequency and phase-speed, we now can give both the definition and the formula for the wave-length. The wave-length λ is by definition the quotient of phase-speed by frequency:

$$(n/2\pi)\lambda = c, \tag{51}$$

and in this special case, the formula for it is:

$$\lambda = 2\pi/|\nabla\varphi|. \tag{52}$$

The reason for giving this quantity a name, and such a name as "wave-length," arises from the best-known and too-exclusively-known special case, that of "plane waves" commonly so called—meaning not only that the wave-fronts are plane, but also that the amplitude is constant over each. Such waves travelling along any direction, say that of x, are described by the expression:

$$f = F \sin (nt - mx), \qquad F = \text{constant}. \tag{53}$$

The wave-fronts—that is to say, the surfaces over which $\varphi = (nt - mx)$ is constant at any moment—are planes normal to the axis of x. These planes are likewise the surfaces over which the displacement f is constant at any moment, and it is tempting to define the wave-fronts as the loci of constant displacement; but this is a coincidence which should be regarded as an accident. At a given moment, any value of $\sin \varphi$

which is found anywhere repeats itself at intervals $2\pi/m$ all along the x-direction; exactly as, at a given point, any value of sin φ which is found at any moment repeats itself at intervals $2\pi/n$ all through time. Owing to the coincidence aforesaid, any value of f which is found anywhere also repeats itself at spacings $2\pi/m$ along the direction of x. This constant spacing serves as the elementary definition of wave-length; and in this special case the elementary agrees with the general definition, for

$$m = |\partial\varphi/\partial x| = |\nabla\varphi| = 2\pi/\lambda. \tag{54}$$

But there is an almost equally simple case in which the spacing between wave-fronts and the spacing between surfaces of constant f are not the same. I refer to the case of spherical waves of sound or the circular ripples on a water-surface, in which we have:

$$f = F \sin (nt - mr), \qquad F = \text{constant}/r. \tag{55}$$

In this case f/F does not conform to the wave-equation, but the function f does. Nevertheless it is the phase, and not the displacement, which advances steadily outward (or inward) in a sequence of steadily diverging (or converging) spherical wave-fronts which expand or contract with the constant phase-speed c. For any two of these spherical wave-fronts differing in radius by $2\pi/m$, the values of ϕ are the same. The surfaces of constant f are also spheres, but they expand or contract with variable speed, and for any two which differ in radius by $2\pi/m$ the values of f are different. This shows that one must not be misled by experience of plane waves into defining "wave-length" as "distance between points where at the same moment the displacement is the same" but must hold fast to the phase as the central fact of any wave-motion.

If the phase-function ϕ does not vary in space, we have the case of *stationary waves*. The coefficient of cos ϕ in the general expression for $\nabla^2 f$ (footnote on p. 339) now vanishes automatically; the coefficient of sin ϕ reduces to the term $\nabla^2 F$, and this must be equated to $-n^2F/c^2$, which if n and c are preassigned leads to an alternative form of the wave-equation

$$\nabla^2 F + \frac{n^2}{c^2} F = 0 \tag{56}$$

very common in acoustics and in wave-mechanics.

In summary:

We have considered two definitions of wave-motion: *first*, that the state of affairs at any point and moment in the medium is controlled by the state of affairs at earlier moments all over any continuous sur-

face drawn in the medium completely around the point; *second*, that the function which is propagated in waves conforms to the so-called "wave-equation."

We have found that these definitions are compatible with one another, the latter being included under the former.

We have applied them to the case of a function which at any particular point of the medium varies as a sine-function of time, thus:

$$f = F(x, y, z) \cdot \sin \varphi; \qquad \varphi = nt - \delta(x, y, z),$$

and have found:

(*a*) that provided the functions F and δ conform to certain stipulations, the function f will satisfy the wave-equation;

(*b*) that φ itself is propagated by wave-fronts; although there is nothing periodic or vibratory about φ, each surface over which φ possesses any constant value wanders onward through space, changing, it may be, in shape as well as position;

(*c*) that the speed with which the wave-fronts of the phase-function φ travel is the speed at which the contributions, out of which the value of f/F at any point and moment is built up, travel to that point from any environing surface.

A Test of Kirchhoff's Theorem

Having formed the conceptions of plane waves and sine-vibrations and frequency and wave-length, we now can practice on Kirchhoff's theorem by applying it to a problem of which the answer is predetermined, so preparing ourselves for other problems of which the answers can be discovered only by means of the theorem.

Imagine plane monochromatic waves, of frequency $\nu = 2\pi n$, wave-length $\lambda = 2\pi/m$, and phase-velocity $c = \nu\lambda = n/m$, travelling in the positive x-direction in an endless procession through an infinite medium. They are described by the function:

$$s = \cos (nt - mx), \tag{61}$$

which is a solution of the wave-equation (6).

The value s_0 of the function s at the origin at the moment t is by hypothesis

$$s_0 = \cos nt \tag{62}$$

and according to Kirchoff's theorem it is given by the following equation:

$$4\pi s_0 = \int dS \left[\cos (n, r) \frac{\partial}{\partial r} \frac{U}{r} - \frac{1}{r} \frac{\partial U}{\partial n} \right], \tag{63}$$

in which the integrand on the right is integrated over any surface completely enclosing the origin. For any such surface, then, the integral on the right of (63) must be equal to $4\pi \cos nt$.

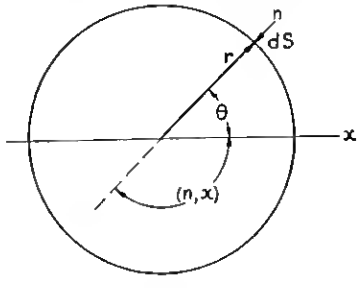

FIG. 1

This we proceed to test for the simplest of such surfaces, a sphere centred at the origin.

Denote by R the radius of the sphere; by θ, the angle between the positive direction of the x-axis and the line drawn from the origin outward through any point P of the sphere. The inward-pointing normal at P is directed oppositely to this line, and hence $\cos(n, r) = -1$ and $\cos(n, x) = -\cos\theta$. The function U and its derivatives are these:

$$\begin{aligned} U &= \cos\left[n\left(t - \frac{r}{c}\right) - mx\right] = \cos(nt - mr - mx), \\ \partial U/\partial r &= m \sin(nt - mr - mx), \\ \partial U/\partial n &= (\partial U/\partial x)\cos(n, x) = (\partial U/\partial x)\cos(\pi - \theta) \\ &= -m\cos\theta \sin(nt - mr - mx). \end{aligned} \tag{64}$$

In each of these we are to set $r = R$ and $x = R\cos\theta$ in preparation for integrating over the sphere. The element of area is

$$dS = 2\pi R^2 \sin\theta d\theta \tag{65}$$

and the limits of integration are 0 and π. Therefore the integral is this:

$$\begin{aligned} I = 2\pi \int d\theta \sin\theta[\cos\varphi - mR(1 - \cos\theta)\sin\varphi]; \\ \varphi = nt - mR(1 + \cos\theta), \end{aligned} \tag{66}$$

which is easy to evaluate, since on putting z for $(1 + \cos\theta)$ and $-dz$ for $\sin\theta d\theta$ it becomes

$$I = -2\pi \int_2^0 dz[\cos(nt - mRz) - mR(2 - z)\sin(nt - mRz)], \tag{67}$$

which can be integrated almost by inspection (the last term through integration-by-parts) and yields the desired value:

$$I = 4\pi \cos nt, \tag{68}$$

and Kirchhoff's theorem comes triumphantly through the test.

It happens however that these conditions, to which we have applied the theorem with so easy a success, are such that the result is of no value whatever in testing the undulatory theory of light.* As I have said already, the eye and the light-recording instruments register amplitude only, not phase. In the train of plane parallel waves described by (61), the amplitude is everywhere the same, and the instruments must report an impression uniformly intense. Nothing could be learned from them about the frequency or the wave-length of the light, and indeed they could not even show that there is anything periodic in the beam. The situation is no better in such a train of spherical diverging waves as (55) describes. Here the amplitude varies inversely with the distance from the centre of divergence, and the eye must report a gradual smooth decline of intensity as it moves away from that centre. In neither observation is there anything to reveal a periodicity or a wave-length, nor anything to forbid the supposition that a beam of light is a stream of straight-flying particles. What we require is a situation in which the use of Kirchhoff's theorem leads to a peculiar and striking variation of amplitude from place to place in the radiation-field—an amplitude-pattern or vibration-pattern, so to speak, depending in detail upon the wave-length of the waves, and so distinctive that if such a pattern of intensity were actually to be found in a field of light one could not but regard it as forceful evidence for the undulatory theory.

Situations which answer this requirement occur when a broad beam of waves is intercepted by a screen pierced with small apertures. In the space beyond the apertures there is a wave-motion of which the amplitude varies from place to place in a remarkable way, depending in detail upon the wave-length. When light falls onto a screen pierced with small holes, the intensity beyond the holes varies remarkably in space. The variations which are observed agree with those which are predicted from the wave-theory, when the proper value of wave-length is chosen; and this is the method of measuring wave-length. Also it is the method of measuring frequency; for the frequency of light cannot be measured directly; it must be computed by dividing the wave-

* In the actual theory of light there are several distinct quantities s—components of electric and magnetic field strengths—each of which separately conforms to equation (6), and which are interconnected in ways which need not yet concern us.

length into the speed of light. Now all the contemporary theories of the atom and of light are based upon values given for frequencies of radiation; and therefore all of them are founded on the wave-theory of light.

We will consequently next consider the propagation of waves beyond apertures, or what Rayleigh called the "effects dependent upon the limitation of a beam of light."

Propagation of a Wave-Motion beyond Apertures

Suppose that the entire plane $x = 0$ is occupied by a wall acting as a total stop to all the waves which come up against it, except where it is pierced by one or more openings; and for simplicity suppose that the oncoming waves compose a plane parallel monochromatic train,

$$s = \cos(nt - mx), \tag{69}$$

advancing from the side of negative x. The primary question is: what goes on in the region to the positive side of the screen?

The question shall be answered by approximations.

The *first approximation* consists in assuming that the wave-fronts come unaltered up to the screen, and each segment which coincides with an aperture goes straight through it and indefinitely onward, travelling unchanged in a straight line even though the surrounding portion of the wave-front has been blocked. If this were perfectly valid, there would be rectilinear propagation of light; the laws of geometrical optics would always be exact; and there would be no need for any but a corpuscular theory of radiation. Because this approximation is deficient, the wave-theory is required. Yet it is close enough to the truth to seem exact to all but the most careful observation, if the apertures are as wide as windows or even as keyholes.

The *second approximation* consists in assuming that the wave-fronts come unaltered up to the screen, and each segment which collides with a portion of the wall is swallowed up and blotted out of existence, but the wave-motion within each aperture is precisely the same as it would be if the entire wave-front passed intact across the plane $x = 0$. That is to say: the displacement on the front face of the screen is supposed to be given thus:

$$\begin{aligned} &s = \cos nt,\ \partial s/\partial x = m \sin nt \text{ wherever there is an aperture,} \\ &s = \partial s/\partial x \quad\ = 0 \qquad\quad \text{wherever there is obstruction.} \end{aligned} \tag{70}$$

This is the assumption on which are founded the conventional theories of the passage of light through a hole, or a slit, or a pair of slits, or a

diffraction-grating, or an echelon, or past a straight-edge or a solid disc or any small obstacle. Having made it, one proceeds to determine the value of s at any point beyond the screen by integrating Kirchoff's integrand all over the apertures—all over the vacant places of the screen, as if in those places only the wave-front were intact, and elsewhere it were abolished; as if the wave-front were cut into the pattern of the apertures as by a template, and each of the segments thenceforth propagated according to the law of wave-motion.

This method is fairly easy to apply, at least when the contours of the openings are simple geometrical figures, circles or rectangles for instance. In practice it is used almost always; and its results are ratified by experience. Yet it is not quite accurate.[4] I am not referring here to the ever-present possibility that boundary-conditions chosen for their mathematical simplicity may not properly describe the actual conditions in the physical world. I am referring to a mathematical, that is to say, a logical, difficulty which is inevitably fatal. A function which is equal to $\cos(nt - mx)$ over arbitrary patches of the plane $x = 0$, but is always equal to zero over the remainder of the plane—such a function does not conform to the wave-equation (6). If in an actual case of light passing through a hole in a screen the phenomena conform everywhere to the wave-equation, then the boundary-condition (70) cannot be valid; if on the other hand the state of affairs in the apertures is rightly described by (70), then the wave-equation cannot be valid and Kirchhoff's theorem cannot be applied.

There is no way out of this dilemma. One must either accept the foregoing assumption frankly as an approximation, or else undertake the vastly more difficult problem of solving the wave-equation itself with the boundary-condition $s = 0$ (or some other which is deemed appropriate) for all the opaque area of the screen, and with other boundary-conditions at infinity to settle the direction from which the light is supposed to come. By this method one makes no assumption about the values of s in the apertures; they are part of the solution. But the general problem is formidable, and no less eminent a man than Sommerfeld was required for the solving of even the simplest conceivable case. Later I will mention his solution of the case in which the barrier covers all that part of the plane $x = 0$ which lies to one side of a straight line, the y-axis for instance, and the remainder of the plane is void. Meanwhile, since on the whole the second approximation is a close one, I will adopt it to explore these effects of *diffraction* which first invited the wave-theory of light, as they now are inviting that of matter; which serve to determine wave-lengths of light, and of matter;

[4] In some of the texts on optics this is not made sufficiently clear.

which set the limits for the powers of telescope and microscope, and perhaps for perception altogether; and which are responsible for the haloes and the parhelia of the sky.

Imagine then that the waves which come up to the screen from behind are plane-parallel and monochromatic, and travel in the positive sense of the x-direction, the screen itself occupying the entire plane $x = 0$. In making the "second approximation" aforesaid, we are to regard this plane as a surface where s and its gradient are zero everywhere save over certain patches—to wit, the apertures—and over these are given by the expressions:

$$\begin{aligned} s &= \cos (nt - mx)_{x=0} \quad = \cos nt, \\ \partial s/\partial x &= m \sin (nt - mx)_{x=0} = m \sin nt. \end{aligned} \tag{71}$$

We are then to determine the value s_0 of s at any field-point P anywhere before the screen—anywhere in the region $x > 0$—by forming Kirchhoff's integral over these apertures:

$$4\pi s_0 = \int dS \left[\cos (n, r) \frac{\partial}{\partial r} \frac{U}{r} - \frac{1}{r} \frac{\partial U}{\partial n} \right]. \tag{72}$$

Over the rest of the plane $x = 0$ the integrand vanishes. Since, however, Kirchhoff's theorem involves an integration over an entire closed surface surrounding P, we ought in strictness to extend the integral over some far-flung surface completing the enclosure; as for instance a hemisphere seated upon the plane $x = 0$, sufficiently great in radius to contain P and all the apertures. This is always neglected, possibly because in practice the wave-motion over such a surface would as a rule be too chaotic to produce any regular effect at P.[5]

In the integrand of (72), r stands for the distance from P to any area-element dS of an aperture; the positive r-direction is measured from P through dS in the direction from front to back; the positive n-direction is the forward-pointing normal to dS, and therefore is identical with the positive x-direction. Remembering the definitions of U and its derivatives, one easily sees that:

$$\begin{aligned} U &= \cos (nt - mr); \\ \partial U/\partial r &= \partial U/\partial x = \partial U/\partial n = m \sin (nt - mr). \end{aligned} \tag{73}$$

It will be convenient to give the symbol θ to the angle between the posi-

[5] Certainly it cannot be argued that the effect from a distant surface is necessarily too small to be noticed at P; we have just seen that in a field of plane-parallel waves it is the same for any spherical surface, no matter how great the radius.

tive x-direction and the line from dS to P, so that $\cos(n, r) = -\cos\theta$. Consequently:

$$4\pi s_0 = \int dS\left[-\frac{1}{r^2}\cos(nt - mr) - \frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right]. \quad (74)$$

One is tempted to say that the quantity under the integral sign is the contribution made by the element-of-wave-front dS to the value of s at P. This notion facilitates both thought and description, and I will adopt it, but with a warning. The danger is that one may come to think of an element-of-wave-front as an independent entity, capable of existing by itself in the medium regardless of what other elements-of-wave-front adjoin it or stand elsewhere. This is unpermissible, for the same reason which makes the method that I am now expounding an approximate and not a rigorous one. Were one of these elements of wave-front alone in the medium, the function s would not conform to the wave-equation. Therefore if we call the expression

$$ds_0 = \frac{1}{4\pi}dS\left[-\frac{1}{r^2}\cos(nt - mr) - \frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right] \quad (75)$$

the *contribution* of the element-of-wave-front dS, we must always remember that it cannot be isolated, but—like the donations to certain endowments—is given only under the condition that other elements also contribute.[6]

The contribution of dS, then, is made up of two terms, one varying inversely as r^2 and the other inversely as r/m. At great distances the latter must increasingly outweigh the former; and "great distances" in this context signify those which are much greater than $1/m$—that is to say, very many times as great as the wave-length of the light. Now as the wave-lengths of most kinds of light are less than .001 mm., a field-point where observations can actually be made must necessarily be distant by many wave-lengths from the screen. Hence it is customary to ignore the first term in the expression (75) and in its integral, and write for the contribution of dS:

$$ds_0 = -\frac{1}{4\pi}dS\frac{m}{r}(1 + \cos\theta)\sin(nt - mr). \quad (76)$$

This expression is the approximate description of what in an earlier

[6] The reader may notice that whereas in dealing with a closed surface surrounding the field-point the n-direction was defined as that of the "inward-pointing normal," there is no way of discriminating between the two senses of the normal to an isolated area-element. This causes an ambiguity in the sign of the contribution; for reversing the sense of the normal reverses the signs both of $\partial U/\partial n$ and of $\cos(n, r)$. The ambiguity is always, I think, physically trivial.

passage I called the "influence" or the "wavelet" which spreads out from the element-of-wave-front in all directions.

Examining it factor by factor, one sees:

(*a*) that the amplitude of the wavelet varies inversely as the distance r from the starting-point, which seems natural;

(*b*) that the wavelet is not isotropic, its amplitude diminishing according to the law $(1 + \cos \theta)$ from a maximum value in the forward to zero in the rearward direction. This is commonly stated as the reason why waves can be propagated in one direction only, not necessarily both forward and backward at the same time;

(*c*) that for waves of the same amplitude and different wave-lengths the amplitudes of the wavelets stand in the inverse ratio of the wave-lengths—the shorter the waves, the more powerfully they are diffracted;

(*d*) that the wavelet from any point is constantly one quarter of a cycle in advance of the primary wave, varying as $-\sin nt$ whereas the wave varies as $\cos nt$.

The advance-in-phase and the factor m in the amplitude enter, it is clear, because the "wavelet" represents the second term in (75)—the term which involves the slope $\partial s/\partial x$ of the wave-function, not the wave-function itself. One might say that the cyclic variation of $\partial s/\partial x$ stirs up a relatively far-reaching commotion in the medium, while the disturbance which the cyclic variation of s excites is rapidly attenuated and mostly negligible. Formerly the factor m and the advance-in-phase seemed unnatural and very strange; for they antedated the theorem of Kirchhoff by sixty years, having been forced upon Fresnel before 1820—and this invites an allusion to history.

Though it is in connection with Huyghens' principle that one commonly hears of wavelets, that principle itself amounts to a denial of nearly every quality which we associate with the ideas of wavelet or wave. Not only are the "wavelets" of Huyghens' construction quite devoid of anything undulatory or periodic; the construction itself is based on the assumption that there is only one point on each where the amplitude is appreciable—the point on the prolongation of the normal from the primary wave-front (corresponding in my notation to $\theta = 0$). But to say that a disturbance is transmitted by wavelets such as these is to say that it is transmitted in concentrated form along lines or rays—which is the same thing as saying that it travels like corpuscles. Huyghens' principle in fact leads straight to the doctrine of the rectilinear propagation of light, and fails either to predict or to explain the phenomena which require a wave-theory.* The accredited

* I am not prepared to say that this is true of the applications to crystal optics.

founder of the wave-theory of light invented in reality a novel language for expressing the corpuscular theory!

Fresnel however invested these wavelets of Huyghens with some of the properties which entitle them to the name. He supposed that the amplitude was distributed widely over each, not confined to the point $\theta = 0$, though greatest at that point; he thought that it diminished slowly with increase of θ, though he did not suggest the precise factor $(1 + \cos \theta)$ nor any other; and he thought that it varied inversely as distance. Further, he endowed it with a periodicity. Thus far, he was right. But naturally he supposed that the cause of the wavelet was the cyclic variation of the wave-function s, and therefore he presumed that it started out in consonance of phase with the primary wave; and he did not insert the factor m. However when he came to test his ideas in somewhat the same way as Kirchhoff's theorem has been tested in these pages—by applying them to a case where the required result was known *a priori*—he was unable to derive the proper answer, except by introducing the factor m and the advance-in-phase; and thenceforth they have figured in the theory of diffraction, indispensable and until the day of Kirchhoff inexplicable.

To return to the problem of determining the wave-motion beyond the apertures: under the approximations stated, it is mathematically quite definite. The solution is the value of the integral:

$$s_0 = -\frac{1}{4\pi}\int dS\left[\frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right], \qquad (77)$$

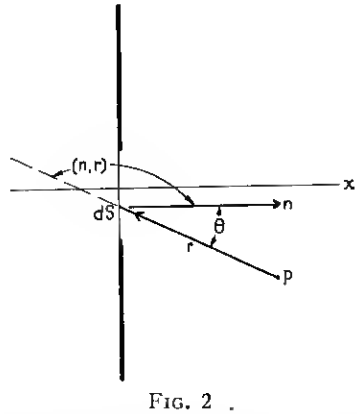

FIG. 2 .

extended over the apertures; r standing for the length of the line joining the field-point P with the element-of-wave-front dS, and θ for the angle between this line and the perpendicular dropped from P to the plane of the screen.

A simple, instructive, and historically famous example is that of the circular hole.

Diffraction from a Circular Aperture

If the propagation of waves were rectilinear, their amplitude would be constant along every line passing normally across an aperture. Such however is as far as possible from being the truth, as we can easily learn by evaluating the wave-motion along the "axis" of a circular hole—that is, the line passing through the centre of the hole perpendicular to the plane of the screen. Locate the centre of the circle at the origin, so that its axis is the axis of x. Denote by R the radius of the circle, by x_0 the coordinate of the field-point P located anywhere upon the axis. All points on any circle centred at the origin being equidistant from P, we may divide the area of the hole by concentric circles into annular elements-of-area. Denote the radius of such a one by p, its breadth by dp; then for it:

$$dS = 2\pi p dp; \qquad r^2 = x_0^2 + p^2; \qquad \cos\theta = x_0/r, \tag{78}$$

and the limits of integration are $p = 0$ and $p = R$.

The problem is now stated in full; but it is very much simplified if the distance x_0 from screen to field-point is very many times as great as the width R of the aperture; and this in practice is commonly the case. Then to first approximation,

$$r = x_0, \qquad \cos\theta = 1, \tag{79}$$

and these values are close enough to the correct ones to suffice for the multipliers of the sine-function in (77); but in the argument of the sine, r is multiplied by m, and a variation of only half a wave-length in r entails a complete reversal of the function; hence in the argument we must proceed to second approximation, and write

$$mr = mx_0 + mp^2/2x_0. \tag{80}$$

Making these substitutions in (77), we have finally:

$$s_0 = -(m/x_0)\int_0^R p \sin(nt - mx - mp^2/2x_0)dp$$

$$= \cos(nt - mx_0) - \cos(nt - mx_0 - mR^2/2x_0) \tag{81 a}$$

$$= \sqrt{2(1 + \cos mR^2/2x_0)}\cos(nt - mx_0 - a). \tag{81 b}$$

Interpreted, these equations tell the startling fact that along the axis of the hole the amplitude, far from being constant, varies in a

gradual and cyclic way between zero as one extreme and double the amplitude of the unintercepted wave-train as the other. As the field-point is displaced along the axis towards or away from the aperture, as the aperture itself is expanded or contracted, doubled agitation succeeds upon quiescence and quiescence upon agitation; and the opening, far from serving as a window to let a segment of the oncoming wave-train pass unaltered by, acts as an agency for producing a curious pattern of varying amplitudes over the region before it.

Now these are precisely the conditions under which, as I remarked before, one can arrange a test of the wave-theory of sound or light; for here we have the amplitude varying from point to point, in a pattern depending in detail upon the wave-length. Experience of light reveals just such a pattern; when parallel light is shed normally upon a screen pierced with a small and accurately rounded hole, the illumination in the axis of the hole passes alternately through maxima and minima as the observer recedes along it. Fresnel was led in a curious way to discover the minima. The French Academy having offered a competitive prize for a study of diffraction—an action instigated, it appears, by adherents of the corpuscular theory of light, who expected that a thorough knowledge of the phenomena of diffraction would demolish the support which they were vaguely supposed to provide for the wave-theory—Fresnel conducted a research and submitted a memoir which ranks among the classics of physical science. It went for judgment to an illustrious committee of five,[7] one of whom, the very eminent mathematician and physicist Poisson—who had been an upholder of the corpuscular theory—promptly deduced the law of the maxima and minima along the axis from Fresnel's conception of the wavelets. He imparted this prediction to the author of the memoir; and in a note appended to the published version, Fresnel has left it on record that he looked for a minimum and found it "like an inkspot" in the centre of the field before the hole.

Equation (81) shows further that the amplitude at any point upon the axis must vary to and fro between the same two extremes—zero, and double the amplitude of the unhindered waves—as the hole expands or shrinks. Wood has described how this may be observed with an iris diaphragm. For an observer stationed at a fixed point upon the axis at a distance x_0 from the hole, the amplitude falls to zero whenever the radius of the circle has one of the values determined by the condition

$$mR^2/2x_0 = \text{even integer multiple of } \pi, \tag{82}$$

[7] Arago, Biot, Gay-Lussac, Laplace and Poisson. It would be hard to assemble a more distinguished group at any time or place.

that is to say, whenever

$$x_0 = mR^2/2k\pi, \qquad k = 0, 2, 4, 6, \ldots, \tag{83}$$

and attains its maximum value, double the amplitude of the uninterrupted waves, whenever

$$x_0 = mR^2/2k\pi, \qquad k = 1, 3, 5, 7, \ldots. \tag{84}$$

Imagine circles drawn upon the plane of the screen, with their common centre at the origin and their radii R_1, R_2, R_3, . . . prescribed by the equations,

$$mR_k^2/2\pi x_0 = k, \qquad k = 0, 1, 2, 3, 4, \ldots. \tag{85}$$

They divide up the plane of the screen into a tiny central circular area and a series of surrounding rings. These are the "Fresnel zones" relative to the point x_0 where the observer is placed. If the circular hole comprises an odd number of the zones, the wave-motion at x_0 attains its maximum; if an even number, the wave-motion vanishes—there is silence or darkness. It seems as if the first, third, fifth and other odd-numbered zones brought light, and the second, fourth and other even-numbered zones destroyed it.

It is equally easy to find the wave-motion along the axis of an annular opening—that is to say, a circular hole partly filled by a concentric circular stop. Denote by R_0 the radius of the stop and by R the radius of the hole; then the limits of integration in (81) are superseded by $p = R_0$ and $p = R$, and the amplitude along the axis varies thus:

$$A = \sqrt{2[1 - \cos m(R^2 - R_0^2)2x_0]}. \tag{86}$$

This contains the surprising conclusion that the maxima of amplitude along the axis are as great as they would be if the stop were removed, though they may be differently placed. An observer properly stationed should see the light brighten when the obstacle is inserted; it may even be brighter than when the obstacle within the hole and the wall surrounding it are totally removed, leaving no hindrance to the onward march of the waves.

The conclusion still holds good when the boundaries of the circular hole retire to infinity, leaving nothing but an opaque disc in an otherwise uninterrupted stream of plane parallel waves; although the approximations made in the foregoing pages are then no longer valid, and equation (86) is not to be employed. Experience however shows that when a small and accurately rounded circular disc is immersed in a beam of parallel light there is a bright spot—more precisely speaking,

a bright core—along the axis of the geometrical shadow. Poisson forecast this also when Fresnel's memoir came before him, and seems to have thought that it would make an *experimentum crucis*, for another member of the committee—Arago—has recorded that he tested the prediction when Poisson made it. He found the bright spot in the centre of the shadow of a circular disc. It is said that Delisle had found and recorded it already, but the record had slipped into oblivion.[8]

We take up now the problem of determining the wave-motion away from the axis—otherwise expressed, that of determining the distribution-of-amplitude over any plane parallel to the plane by the screen.

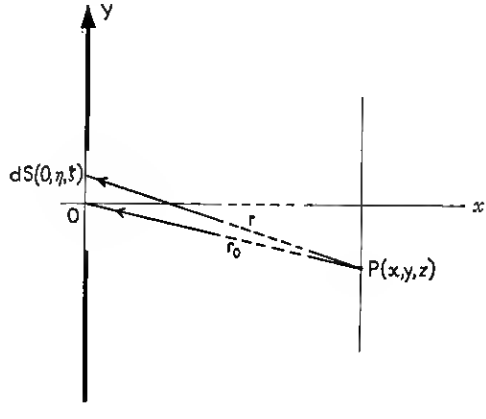

FIG. 3

Denote by (x, y, z) the coordinates of any field-point and by $(0, \eta, \zeta)$ those of any area-element dS of the aperture; by r, as heretofore, the distance from P to dS, and by r_0 the distance from P to the origin. Then

$$r^2 = x^2 + (y - \eta)^2 + (z - \zeta)^2 = r_0^2 - 2y\eta - 2z\zeta + \eta^2 + \xi^2. \quad (87)$$

As heretofore r and r_0 shall be supposed to be very many times as great as the dimensions of the apertures, and therefore as the greatest values attained by η and ζ; therefore, to first approximation,

$$r = r_0, \qquad \cos\theta = x/r_0, \quad (88)$$

[8] It is interesting to notice why an accurately circular disc is required to show the bright spot in its best development. Take the case of the aperture, since we already have its suitable equation (81). A nearly but not quite circular hole may be regarded as made up of sectors, each with a different radius. For each of these the upper limit of the integral in (81) would be different, and therefore the condition (84) for doubled amplitude could not be realized for all at once. There would be a wave-motion along the axis, but not the regular alternation of maxima and minima nor the sharply outstanding brightness at the maxima.

and to second approximation,

$$r = r_0 - (y\eta + z\zeta)/r_0. \tag{89}$$

Into equation (77) we insert the first-approximation values of r and $\cos\theta$ in the multipliers of the sine-function, but the second-approximation value of r into the argument of the sine. Therefore we have, for the value of s at the very distant point (x, y, z), the expression:

$$s = -\frac{1}{4\pi}\frac{m}{r_0}\left(1 + \frac{x}{r_0}\right)\int\int d\eta d\zeta \sin\left(nt - mr_0 - m\overline{y\eta + z\zeta}/r_0\right). \tag{91}$$

It is expedient to introduce the three direction cosines of the line extending from the origin to the field-point, the cosines of the angles between it and the coordinate axes:

$$\alpha = \cos(x, r_0) = = x/r_0; \qquad \beta = y/r_0; \qquad \gamma = z/r_0. \tag{92}$$

Then, with a slight additional transformation, we convert equation (89) into:

$$\begin{aligned} s = \text{const.}\,(1+\alpha)\Big[&\sin(nt - mr_0)\int\int d\eta d\zeta \cos m(\beta\eta + \gamma\zeta) \\ &- \cos(nt - mr_0)\int\int d\eta d\zeta \sin m(\beta\eta + \gamma\zeta)\Big] \\ = \text{const.}\,(1+\alpha)&[C\sin(nt - mr_0) - S\cos(nt - mr_0)], \end{aligned} \tag{93}$$

the symbols C and S being traditional for these integrals.

The coordinates of the field-point have disappeared, leaving only the cosines which define its direction as seen from the origin. This means that we have here the formula for the wave-motion over any plane parallel to the screen and infinitely far away, in terms of the directions in which its various points are seen. The words "infinitely far away" sound formidable; but it is not necessary to depart for infinity, in order to find a plane where (93) describes the state of affairs. There is an artifice for bringing the infinitely distant plane up to a convenient nearness; an artifice known as a *lens*. When a converging lens is set up before the apertures, the wave-motion predicted by the formula (93) for all points infinitely far away upon the line with direction cosines (α, β, γ)—this wave-motion occurs at the point where the line intersects the focal plane of the lens. Therefore we may regard equation (93) as the description, according to the wave-theory of light, of the distribution-of-amplitude in the focal plane of the lens. (To convert the cosines into coordinates in that plane, it is sufficient to multiply each by the focal length of the lens.)

Returning now to (93), it is evident that the problem is solved when the integrals are evaluated; in particular the amplitude is given by the formula,

$$A = \text{const.}\ (1 + \alpha)\ \sqrt{C^2 + S^2}. \tag{94}$$

Whatever the shape of the aperture or apertures, the values of the integrals can be determined as closely as may be desired; and in two instances which happily are the most frequent and useful—those of the circular and the rectangular openings—the integrations lead directly to familiar functions.

Diffraction Patterns in the Focal Plane of a Lens

If the origin is located at the centre of the circle, the integral S vanishes—for the value of the sine-function contributed by each area-element is annulled by the value contributed by the element symmetrically placed to the other side of the centre—and the integral C for the same reason becomes this:

$$C = \int\int \cos\ (m\beta\eta)\ \cos\ (m\gamma\zeta) d\eta d\zeta. \tag{95}$$

By putting $\gamma = 0$ and then integrating, we shall obtain the distribution of amplitude along the line passing through the centre of the diffraction-pattern and parallel to the axis of y; but this, by reason of the circular symmetry of the entire system, is the same as the distribution of amplitude along any radius passing through the centre of the diffraction-pattern, and therefore is all we need. In the expression so obtained, replace the Cartesian coordinates heretofore used in the plane of the screen by polar coordinates ρ and φ; then we have

$$C = \int\int \rho \cos\ (m\beta\rho \cos\ \varphi) d\rho d\varphi, \tag{96}$$

the limits of integration being 0 and R (the radius of the aperture) for ρ, and 0 and 2π for φ.

The integral C is proportional to the Bessel function of order unity of the variable $mR\beta$:

$$C = 2(\pi R/m\beta) J_1(mR\beta). \tag{97}$$

This is a function which like the sine vanishes at intervals, though not at equal intervals. The centre of the diffraction-pattern is therefore encircled by concentric rings over each of which the wave-motion vanishes; between each pair of these there is a zone where the amplitude differs from zero and varies, attaining a maximum somewhere near the middle of the zone. In the focal plane of the lens there are ring-shaped zones of light, surrounded and divided by dark circles; these are the "fringes."

This system of annular fringes is the *image* produced by a lens on which plane-parallel light falls normally through a circular aperture; also when there is no screen before the lens, for being circular it serves as its own aperture. Now plane-parallel light is such as originates in an infinitely distant luminous point, or—what comes to the same thing —any luminous object so distant that neither the curvatures of the wave-fronts proceeding from its various parts nor the angles between the directions in which these lie are appreciably large; a star, for instance. The image of a star in the focal plane of a telescope objective is therefore not a point, however far away the star may be; it is a system of rings. So it is in the eye, the pupil serving as the aperture; but the inner rings are in both cases so narrow and the outer rings so faint that they appear condensed into a point. Magnification of the fringes in the telescope by the eyepiece brings them into view, and so they set a limit to the value of magnification; for it is of no avail to be able to examine an image more minutely if all that can be examined is the consequence of the disturbance produced in the incoming waves by the finiteness of the lens.

The limitation which the law of propagation of light thus sets upon the formation of images is very important. The simplest possible illustration is furnished by a double star. Let the telescope be directed upon such a pair of stars so that the light from one component falls normally upon the lens, the light from the other component at any angle of which I denote the complement by φ; thus φ stands for the angular distance between the two stars in the sky. Now I have not hitherto treated the case of light falling otherwise than normally upon the screen containing the apertures, which in this case is nothing but the plane of the objective. The extension however is immediate. Orienting the y-axis in the plane of the screen so that the direction of propagation of the waves coming from the stars lies in the xy-plane, we have for the wave-function in the region extending up to the aperture from behind:

$$s = \cos(nt - mx\cos\varphi - my\sin\varphi), \tag{98}$$

and therefore in the plane of the aperture ($x = 0$) we have, instead of the values given in (71), these:

$$\begin{aligned} s &= \cos(nt - my\sin\varphi), \\ \partial s/\partial x &= m\cos\varphi\sin(nt - my\sin\varphi), \end{aligned} \tag{99}$$

and for the value of s at any point in front of the aperture we have, instead of (77), this value:

$$s_0 = -\frac{1}{4\pi}\int dS\left[\frac{m}{r}(\cos\theta + \cos\varphi)\sin(nt - my\sin\varphi - mr)\right],$$

and there are corresponding changes in the values of the integrals C and S which determine the amplitude. To first approximation—that is to say, when φ is not too great—the result is, that the diffraction pattern of one star is like that of the other, but shifted sidewise. The angular displacement between the centres of the two fringe-systems is the same as the angular displacement between the two stars. The question now arises: how far apart must the two fringe-centres be, that the two families of rays may be securely told apart?

Such a question of course cannot be definitely answered; the answer would depend upon the acumen and the experience of the observer. The conventional response is, that the two systems of rings are surely distinguishable if the centre of one lies upon the first dark circle of the other. Now the angular radius of the first dark ring, i.e., the value of β for which the Bessel function of (97) first vanishes, is 1.22 times the ratio of the wave-length of the light to the diameter of the aperture; for green light in the largest available refracting telescope this amounts to about an eighth of a second of arc. This then is nearly the least angular separation between two stars which are distinguishable; a pair or a group much closer together would appear as one, not through any avoidable defect of the telescope nor through any insufficiency of the eyepiece but through the laws of propagation of light themselves, working to prevent the formation of an image indefinitely sharp.

For a rectangular aperture the integrals C and S are extremely easy to evaluate. The diffraction-pattern is a criss-cross of dark lines, intersecting at right angles and bounding rectangular areas of light, similar in shape to the aperture but oriented at right angles to it. If the rectangle is prolonged indefinitely and so becomes an infinitely long slit, the diffraction-pattern becomes a sequence of parallel bands separated by dark lines normal to the length of the slit. If then a multitude of identical slits are cut into the screen at equal intervals side by side, a new periodicity is superposed upon the periodicity of the waves, and out of the interaction of these two there come diffraction-patterns much more sharp and striking than any which a single aperture, however shaped, is able to produce. These will be considered in the following chapter.